Research traditions and emerging expectations: PhD students and their research data management

Kerstin Helbig

Computer and Media Service, Humboldt-Universität zu Berlin, Germany

kerstin.helbig@cms.hu-berlin.de

Abstract

Research data management has become an important topic for research institutions around the world. More and more funders demand transparent research as well as data management plans. In accordance, universities start implementing data policies to further proper data handling. Humboldt-Universität zu Berlin wanted to shed light on doctoral candidates' current research data management. Therefore a short survey was conducted in 2015 with 187 participants. Results show that a majority of the respondents uses digital research data for their dissertations. Of those doctoral candidates only a minority has a concept for archiving their research data. Even though preservation is unclear, many PhD students with digital data plan to make it open. In compliance with these results a subsequent workshop for doctoral candidates was organized. Especially PhD students from the humanities expressed difficulties to name their produced or analyzed research data. They also mentioned legal concerns more often than other doctoral candidates. Students from the sciences had a better understanding of research data management. Reservations against data sharing were rare among them.

Keywords

research data management; training; doctoral candidates; information literacy

Introduction

Higher education institutions and research funders place new demands on today's researchers. Especially doctoral candidates and early-career researchers face changing research traditions as new policies and funding requirements emerge. Funder guidelines from German Research Foundation (Deutsche Forschungsgemeinschaft, 2015) or European Commission (2016) and institutional policies like the research data management policy from Humboldt-Universität zu Berlin (2014) establish standards of good scientific practice regarding research data. The aim is an easier replication and verification of research results as well as an acceleration of research and a better transfer of basic research results to industry. PhD students are confronted with these new expectations. However, traditional research methods and procedures of scientific communication most often rule in practice and can hinder adoption.

English speaking countries like the United States of America, Australia or the United Kingdom have much longer experience in data curation, data literacy, and information dissemination. National competence centers such as the British Digital Curation Center (DCC)[1] or the Australian National Data Service (ANDS)[2] reflect the structural embeddedness of the topic. Germany and other continental European countries are just at the beginning of structured and comprehensive research data management. In Germany, nationwide research data management is not yet fully established, though more and more initiatives emerge at German higher education institutions. Research data management initiatives from the very start were among others University of Bielefeld, Technische Universität Berlin, Heidelberg University, and Ludwig-Maximilians-Universität München. First actions of these initiatives included in various composition the establishment of an open science bureau, the adoption of a research data policy,

[1] Digital Curation Center http://www.dcc.ac.uk (accessed 2 June 2016)

[2] Australian National Data Service http://www.ands.org.au (accessed 2 June 2016)

the set-up of an institutional data repository, and the launch of training and information sessions. Frequently, the universities' data center and library as well as other partners widely collaborate to establish the best service possible. More than a few German universities are, however, still in the orientation phase and work on a strategic approach.

In 2012, the research data management initiative of Humboldt-Universität zu Berlin was established by the Vice President for Research. The three central units Computer and Media Service, Research Service Centre, and University Library collaborate to enhance attention and improve data management. Initial efforts of the initiative included the adoption of a research data management policy (Humboldt-Universität zu Berlin, 2014), the conducting of a survey (Simukovic et al., 2013) and face-to-face semi-structured interviews about research data management among the university's researchers (Simukovic et al., 2014). Empirical results substantiated the need for support and guidance. In 2015, the initiative thus began to offer training and support for research data management and launched workshops (see e.g. Helbig, 2016). This included the start of a collaboration between the research data management initiative and the Humboldt Graduate School. The collaboration intends to improve data literacy (Jones et al., 2013: 17) as well as good scientific practice among doctoral candidates.

This paper summarizes outcomes from a small survey among PhD students. Furthermore, a workshop approach is described that outlines how support and improvement of research data management among young researchers can be initiated. Problems and obstacles of doctoral candidates are outlined. Experiences and recommendations from our collaboration project with the graduate school complete the paper.

Methodology

Through a short survey Humboldt-Universität zu Berlin wanted to assess the *status quo* of doctoral candidates and their research data management. The online questionnaire was available for about two weeks from 19 January to 31 January 2015 (Kindling, 2016). The survey was announced through Humboldt Graduate School, central mailing lists, and the initiatives' research data management website[3]. The questionnaire comprised eight questions, both multiple-choice and open. The basic population of doctoral candidates at Humboldt-Universität zu Berlin cannot be determined as a formal registration is often made only shortly before submission. Consequently, the survey cannot be seen as representative. However, the results can provide a first idea about data management among PhD students. LimeSurvey as well as SPSS were used as software for data collection and statistical analysis, respectively.

The workshop conception and implementation was based on our discipline-specific workshop approach (see Helbig, 2016). Necessary adjustments included an increased focus on requirements of funders and of the university outlined by their respective policies. Data citation and the use of licenses were emphasized as well to cater audience-specific needs. Answers of participants during the brainstorming were first structured, stemmed, and categorized. Visualization was then realized using Wordle.

Results

Survey

187 doctoral candidates took part in the survey. As the population is unknown, results need to be seen as non-representative. Furthermore, the distribution is not even across disciplines (life

[3] Research data management website of Humboldt-Universität zu Berlin http://hu.berlin/dataman (accessed 2 June 2016)

sciences 7.5%, humanities and social sciences 76.5%, natural sciences 13.4%, other 2.7%[4]). This skewed distribution partially results from the mailing lists that were used to advertise the survey. Nevertheless, first deductions can be made.

A majority of the respondents uses digital research data for their dissertations (see Figure 1; 71.7%). Of those doctoral candidates only a minority has a concept for archiving their research data (21.6%). Consequently, there is a need for support and information about storage and long-term archiving expressed by a major part of doctoral candidates. Even though preservation is unclear, many PhD students with digital data plan to provide public access to their research data (43.3%). Results suggest that a considerable number of doctoral candidates want to be transparent about their research and are willing to share their data with fellow researchers.

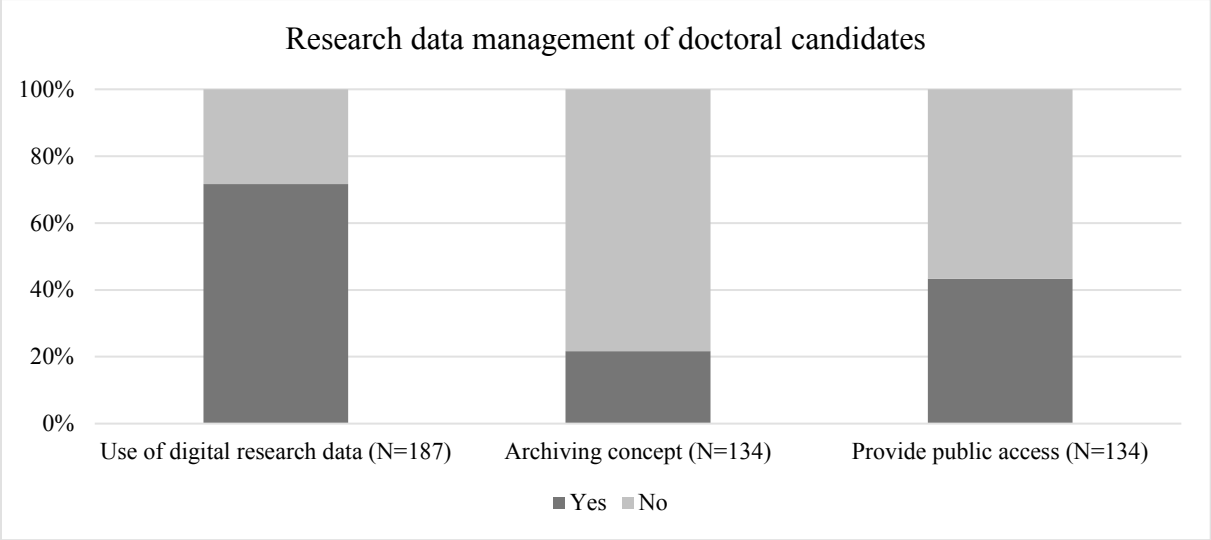

Figure 1: Research data management of doctoral candidates at Humboldt-Universität zu Berlin (Source: own calculation)

Research data management training within the doctoral program is therefore needed to improve data literacy and the management of research data. In addition, technical solutions need to be offered to allow for easy data sharing and preservation of dissertation data. Following these empirical results, the research data management initiative created and organized a workshop for doctoral candidates in cooperation with Humboldt Graduate School.

Research data management workshop

In compliance with these empirical outcomes a workshop with doctoral candidates was organized in winter semester 2015/16. The training was prepared and advertised in cooperation with Humboldt Graduate School. The initiatives' research data management website, social media, and bulletin boards at the University Library informed inter alia about the training. The three trainers encompassed the research data management coordinator of the university and two subject librarians of the University Library who specialize on information and data literacy.

The curriculum was divided into a general introduction into research data management and a separate discipline-specific group session. The conception provided for 120 minutes duration – 60 minutes general introduction and another hour for group work and question-and-answer session. After a short brainstorming about research data, the group session addressed relevant repositories, file formats, and metadata standards. The following content was presented.

[4] The classification is based on the subject taxonomy of the German Research Foundation (2012-2015).

- General introduction into research data management
    - Definition of research data and its management
    - Policies and requirements of funders and the university
    - Services offered by our initiative
    - Data management planning
    - Backup and security of data
    - File structuring
    - Long-term preservation
    - Choice of licenses and data citation
- Discipline-specific group session
    - Brainstorming about own research data as well as benefits and possible reservations regarding a transparent research process
    - Recommendable file formats
    - Documentation and relevant metadata standards as well as contact persons
    - Presentation of relevant discipline-specific repositories and archives
    - Discipline-specific research data journals
- Joint discussion and question-and-answer session

Target group were doctoral candidates from all disciplines and in all years of PhD studies. The reception was outstanding and emphasizes the interest as well as the need for guidance. More than thirty doctoral candidates registered for the workshop. Eleven persons were on the waiting list. Mainly PhD students from biology, medicine, physics and agriculture took part. Though, doctoral candidates from the humanities attended, as well.

For the group session the PhD students were divided into discipline-specific groups of about 10 persons each: biology and medicine; agricultural sciences, biology and physics and the last consisting of students from different backgrounds like social sciences, economics, psychology, humanities, rehabilitation sciences, archeology and law. The brainstorming gave important information about the PhD students' research data and obstacles with research data management and sharing (see Figure 3). Data types mentioned most often across disciplines include videos and other visual data (photos, images, pictures) as well as code. Formulated benefits encompass fast and easy access, the opportunity of data reuse, and reproducibility. Legal and copyright issues, but also concerns of plagiarism were named most frequently as reservations against research data sharing.

Figure 3: Word cloud of answers made by doctoral candidates during the group session (Source: own calculation)

Discipline-specific differences were though apparent. During the group session, especially PhD students from the humanities expressed difficulties to name their produced or analyzed research data. Some were uncertain what digital research data means for them. They also mentioned legal concerns more often than other doctoral candidates. Copyright and privacy issues were stated in particular. Restricted access can hence be inevitable to enable data sharing at all, especially in the social sciences or medicine. In addition, legal grounds regarding research data have to be clarified. In contrast to PhD students from the sciences, social scientists and humanists were very much aware of the expenditure of time. As data documentation and sharing can be time-consuming, it is necessary to take that into account in the dissertation process.

Doctoral candidates from the sciences had a better understanding of research data management. They were able to name their digital research data in detail (e.g. file format of sequencing data). Reservations against data sharing were rare among them. Cooperation and the impact on future research played a larger role. Not only positive, but also negative aspects of research data sharing were though mentioned. Misuse, fraud and plagiarism were perceived as the most eminent problems of open research data. Embargo periods are consequently more of interest for scientists.

Besides the brainstorming, doctoral candidates added benefits and obstacles during the final discussion, as well. It became obvious that some PhD students were quite experienced in data sharing and open science. Others had little knowledge in regard to data management. The level of knowledge is hence diverse. Noted obstacles also encompass research traditions that are imposed by fellow researchers or supervisors. As data sharing is not common in all disciplines, conflicts can arise. Open discussion or conflict consultation, which is offered at Humboldt Graduate School, can be a solution.

Discussion

Specifics of PhD research data management

Taking all results into account, some specifics of doctoral research data management can be summarized. Even more than other researchers, PhD students are under enormous time pressure due to a more or less fixed dissertation period. Therefore research data management should be integrated early in the dissertation process. Training should ideally be scheduled in the beginning of the dissertation and pursued throughout the whole dissertation process, complementary to the research data lifecycle[5].

On the personal level some factors can interfere with data sharing and management. Emerging funder policies might contradict the traditional publication process that is practiced by fellow researchers or supervisors. This can result in a dilemma for PhD students. On the one hand they want to meet discipline-specific expectations. On the other hand doctoral candidates have to comply with the requirements of their funders. Furthermore, distrust regarding fellow researchers and industry have to be taken into account, especially in the sciences. Embargo periods but also discussion within the scholarly community can dispel such concerns.

Furthermore, doctoral candidates are very much aware of publication pressure. As long as traditional publishing methods are of greater significance than other forms of publication, incentives are missing. Credit needs to be given to enhance the appeal of sharing research data (Fecher et al., 2015). In addition, appointment procedures need to change in favor of open science publications.

Valuable insights were also gained on the institutional level. Dissertation projects result in a variety of data formats that may have to be handled within the institutional research data

[5] http://www.data-archive.ac.uk/create-manage/life-cycle (accessed 2 June 2016)

repository. The diverse knowledge base of the doctoral candidates is a challenge for training measures. Legal expertise is expected, thus requiring the consult of experts to competently provide information. The current unclear legal situation in German copyright and few contact points however complicate consultancy.

New training offerings and technical solutions raise awareness and assist researchers. Discipline-specific training is needed to support doctoral candidates and foster data literacy. Above all practical information is desired by the PhD students. To meet this demand, our initiative offers e.g. a workshop on data management plans that teaches skills for the independent creation of a plan. Participants are introduced to the tool DMPonline of the British Digital Curation Centre that assists with plan creation. This makes data management planning faster and easier for the doctoral candidates.

Recommendations for other higher education institutions

- Cooperate with strategic partners like graduate schools or doctoral networks
- Offer technical solutions for secure storage, file exchange, file management, version control, password management, and data publication
- Include research data into electronic thesis consultation
- Start with research data management training as early as possible during the dissertation process
- Continuously offer advice and information material
- Try to respond to discipline-specific needs
- Raise awareness of dissertation supervisors through information events
- Offer consultation for conflict situations
- Advocate the recognition of a transparent research process at the institution
- Create guidelines and standards for research data management within the institution

Conclusions

The experiences from the workshop confirm our initial empirical survey results. The workshop feedback showed that not only theoretical guidance and information material are of importance. PhD students desire support in their daily research data management. Practical tools and software solutions need to be offered. Actively supporting research tools like digital laboratory journals, annotation and file sharing software have to be secure, easy to use as well as legally unproblematic. In addition to analogue training for data literacy, the initiative is currently developing online tutorials which will teach researchers and other interested parties without restriction on time and location on the subject of research data. Since only tutorials in English exist at the moment, we see the need to offer these tutorials in German. An English translation is planned for later on in the process as well. Especially discipline-specific information videos will expand the range of tutorials. The goal is to reach not only doctoral candidates, but also other members of Humboldt Universität zu Berlin, such as librarians, research fellows, and professors. To improve data skills, it takes support before, during and after the dissertation. Hence, more workshops and technical changes in the context of the project eDissPlus, funded by the German Research Foundation (DFG), are underway.

Acknowledgements

The author thanks Pamela Aust and Ulrike Schenk for their invaluable support and assistance with the workshop. Maxi Kindling designed the survey questionnaire and provided the data for statistical analysis. Final thanks go to the doctoral candidates for their participation and comments.

References

Deutsche Forschungsgemeinschaft (2015). *DFG Guidelines on the Handling of Research Data*. Available at http://dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/guidelines_research_data.pdf (accessed 2 June 2016).

European Commission (2016). *Guidelines on Data Management in Horizon 2020. Version 2.1*. Available at http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf (accessed 2 June 2016).

B. Fecher, et al. (2015). 'A Reputation Economy: Results from an Empirical Survey on Academic Data Sharing'. *DIW Berlin Discussion Paper* 1454: 1-28. doi:10.2139/ssrn.2568693

K. Helbig (2016). 'Research Data Management Training for Geographers: First Impressions'. *ISPRS International Journal of Geo-Information* 5(4): 40. doi:10.3390/ijgi5040040

Humboldt-Universität zu Berlin (2014). *Humboldt-Universität zu Berlin Research Data Management Policy*. Humboldt-Universität zu Berlin, Berlin. Available at https://www.cms.hu-berlin.de/de/ueberblick/projekte/dataman/hu-rdm-policy/view (accessed 2 June 2016).

S. Jones, et al. (2013). *Research Data Management for Librarians*. Digital Curation Centre, Edinburgh. Available at http://www.dcc.ac.uk/sites/default/files/documents/events/RDM-for-librarians/RDM-for-librarians-booklet.pdf (accessed 2 June 2016).

M. Kindling (2016). 'Research Data Management at Humboldt-Universität zu Berlin – Status Quo and Perspectives'. In *Séminaire DRTD-SHS « Les données de la recherche dans les humanités numériques » February 2nd, 2015,* available at http://de.slideshare.net/MaxiKindling/research-data-management-at-hu-berlin-status-quo-and-perspectives (accessed 2 June 2016).

E. Simukovic, et al. (2013). *Umfrage zum Umgang mit digitalen Forschungsdaten an der Humboldt-Universität zu Berlin*. Humboldt-Universität zu Berlin, Berlin. urn:nbn:de:kobv:11-100213001

E. Simukovic, et al. (2014). *Was sind Ihre Forschungsdaten? Interviews mit Wissenschaftlern der Humboldt-Universität zu Berlin*. Humboldt-Universität zu Berlin, Berlin. urn:nbn:de:kobv:11-100224755

Biography

Kerstin Helbig is research data management coordinator at Humboldt-Universität zu Berlin. In her consultative capacity, she assists researchers in the management of their research data and organizes training as well as information sessions. In her former position she was a research associate at GESIS - Leibniz Institute for the Social Sciences. In the da|ra project – an allocation agency for Digital Object Identifiers (DOI) in Germany funded by the German Research Foundation (DFG) – she was responsible for the further development of the used metadata schema. In addition, she supported researchers in the registration of their research data. Through her study of social sciences and many years of experience at GESIS, she has profound knowledge in the handling of research data.